Application of Attention Mechanism with Bidirectional Long Short-Term Memory (BiLSTM) and CNN for Human Conflict Detection using Computer Vision

Erick da Silva Farias[1], Eduardo Palhares Júnior[2], Nivaldo Rodrigues e Silva[3], Alexandre Lopes Martiniano[4]

[1,2,3,4] Federal Institute of Amazonas, Manaus 69083-000, Brazil
edsfrlinux@gmail.com, eduardo.palharesjr@ifam.edu.br, nivaldo@ifam.edu.br, alexandre.martiniano@ifam.edu.br

Abstract

The automatic detection of human conflicts through videos is a crucial area in computer vision, with significant applications in monitoring and public safety policies. However, the scarcity of public datasets and the complexity of human interactions make this task challenging. This study investigates the integration of advanced deep learning techniques, including Attention Mechanism, Convolutional Neural Networks (CNNs), and Bidirectional Long Short-Term Memory (BiLSTM), to improve the detection of violent behaviors in videos. The research explores how the use of the attention mechanism can help focus on the most relevant parts of the video, enhancing the accuracy and robustness of the model. The experiments indicate that the combination of CNNs with BiLSTM and the attention mechanism provides a promising solution for conflict monitoring, offering insights into the effectiveness of different strategies. This work opens new possibilities for the development of automated surveillance systems that can operate more efficiently in real-time detection of violent events.

Keywords: *Deep Learning, BiLSTM, CNN, Attention Mechanism*

1. Introduction

Violence is a complex phenomenon that permeates the history of humanity, manifesting itself in different ways and in different contexts. Since the beginning of civilization, violence has been present in wars, territorial conflicts, and power disputes. Over time, new manifestations emerged, such as domestic violence, urban crime, and terrorist attacks. Violence manifests itself in seemingly trivial situations, such as fights in bars or traffic conflicts. These episodes reflect social tensions, accumulated frustrations, and, often, the lack of adequate conflict resolution mechanisms. The culture of aggression and the normalization of violence in social relationships can intensify these situations, creating a cycle that is difficult to break. In many contexts, social inequality, poverty, and marginalization also fuel violence, creating an environment conducive to organized crime and urban violence. Thus, violence in today's world is a multifaceted phenomenon that requires a critical and multidisciplinary approach to understand it and, above all, combat it. Analysis of its historical, social, and cultural roots is fundamental to developing effective prevention and intervention strategies. Surveillance cameras are widely used in commercial establishments, homes, industries, schools, and public places. These cameras are intended to assist agents who monitor the location, however, this type of conventional monitoring is not very effective when hundreds of cameras are deployed because of human involvement, because identifying incidents using conventional cameras becomes an inefficient task. An efficient way to identify incidents via a surveillance camera would be through computer vision, because images from the CCTV system can be linked to a trained deep learning model to make inferences about incidents related to violence between humans in real time. This approach to using computer vision is relevant as it will eliminate the cost of surveillance by humans. But for this to work, it is necessary to carry out tests, collect images to train the model, compare deep learning models, and other adjustment processes to refine the human conflict detection system.

With respect to data collection, it is important that the data set has a significant volume, with variance in class data and good resolution. According to (Dashdamirov, 2024), for effective algorithm training, the collection and labeling of a vast volume of data is essential. Although there are public sets of videos available, there is still a significant need to expand the amount of this data. Furthermore, aspects such as video resolution, frame frequency, lighting conditions, and camera angles vary greatly. These differences complicate the development of models that are both robust and capable of generalizing appropriately. The use of Deep Learning in the context of human conflict monitoring is relatively new, because the data available publicly has a small volume and has low quality in the video frames. (Dashdamirov, 2024) evaluates deep learning techniques in detecting violence in videos, highlighting that increasing the dataset from 500 to 1,600 videos improves the average accuracy of the models by 6%. It demonstrates the importance of large data sets and transfer learning for more effective surveillance systems. (Datta et al., 2002) analyzed the trajectory of movements and orientation of body limbs to

detect violent behavior. (Nguyen et al., 2005) introduced a hierarchical hidden Markov model (HHMM), showing that it can be useful for recognizing aggressive attitudes, especially through a standard HHMM approach aimed at identifying violence. (Kim & Grauman, 2009) combined probabilistic Principal Component Analysis (PCA), used to identify flow patterns in local areas, with Markov Random Fields (MRF), which help maintain global model coherence. On the other hand, (Mahadevan et al., 2010) argued that optical flow-based representations are not suitable for detecting unusual changes in both appearance and motion. They proposed a technique that identifies violent scenes by evaluating elements such as the presence of blood, flames, intensity of movement and sound volume.

2. Methodology

In this chapter the methodology will be presented. In 2.1, computer vision was discussed. In section 2.2 Deep Learning and Neural Networks were covered, LSTM and BiLSTM in the subtopics and in session 2.3 about the Attention Mechanism.

2.1 Computer Vision

Computer vision is an area of artificial intelligence (AI) that deals with developing methods that allow computers to acquire, process and interpret visual information from the real world, with the aim of making decisions or providing recommendations (Szeliski, 2010). The main challenges of computer vision in videos involve the need to identify and classify objects and actions in dynamic environments, such as recognizing human behavior patterns or detecting specific events, such as conflicts, aggressions or complex interactions.

2.2 Deep Learning and Neural Networks

Deep learning is a subfield of artificial intelligence that relies on deep neural networks to perform complex recognition, classification, and prediction tasks. These networks are composed of multiple layers of processing, allowing them to learn hierarchical representations of data such as images, text and temporal sequences.

Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) are a class of deep neural networks that have been widely used in computer vision tasks due to their ability to learn efficient representations of visual data. Figure 1 shows a diagram that represents the architecture of a CNN. They are composed of convolutional layers, pooling layers, and fully connected layers. CNNs are effective in extracting spatial features from images, which allows them to detect patterns, such as edges, textures and shapes (LeCun et al., 2015).

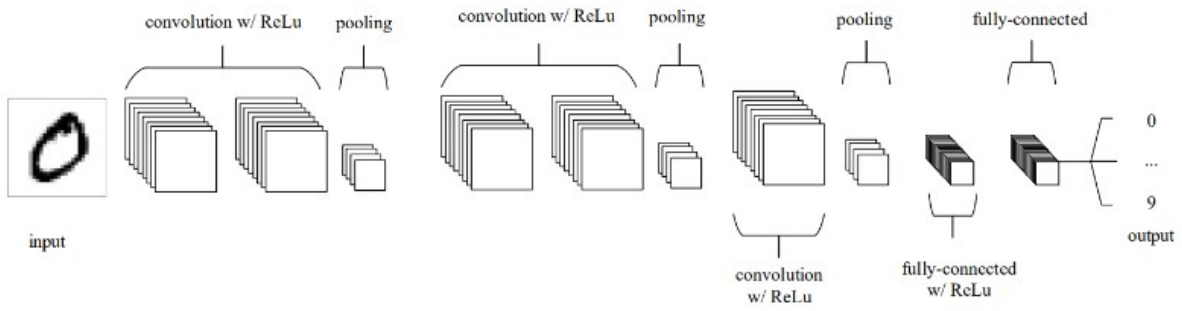


Figure 1. An example of CNN architecture (O'Shea & Nash, 2015).

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a recurrent neural network (RNN) architecture designed to model temporal dependencies in sequential data. Figure 2 shows the LSTM architecture diagram. LSTMs have memory cells that allow the retention of information over time, overcoming the problem of gradient fading that limits other traditional RNNs (Hochreiter & Schmidhuber, 1997).

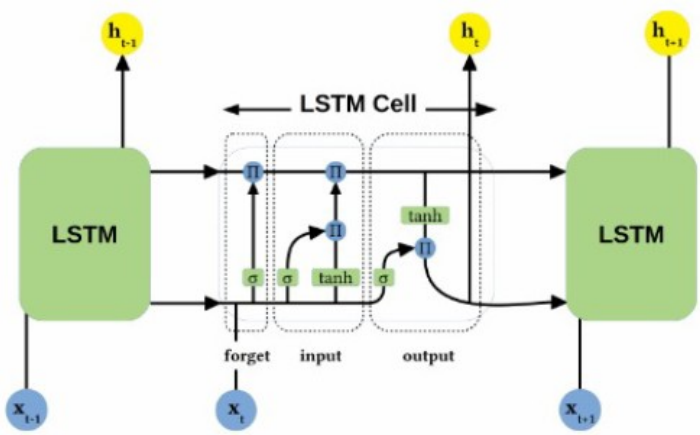


Figure 2. Basic Long-Short Term Memory (LSTM) architecture [Shenfield and Howarth 2020].

2.3 Attention Mechanism

The Attention Engine is a fundamental technique in the field of deep learning, used to improve the ability of models to focus on the most relevant parts of input during processing. Instead of treating all elements of the input equally, the Attention Engine allows the model to learn to allocate greater weight to the most informative parts of the input, improving the efficiency and accuracy of predictions. This technique was initially proposed by (Bahdanau et al., 2015), and later refined into models such as the Transformer proposed by (Vaswani et al., 2017), which use attention mechanisms as their central basis. In Weighted Sum Attention Mechanism, the model calculates an attention score for each element of the input sequence, using a dot product between the input vector and a weight vector learned during training.

Designed Architecture for Conflict Detection

The model developed for classifying image sequences is based on a deep neural network designed to capture both spatial and temporal features from the data. The input consists of a sequence of 15 images, each with a size of 100x100 pixels and 3 color channels (RGB), initially processed by a convolutional layer (CNN). The first processing step applies a TimeDistributed layer, allowing the convolutional network to treat each image independently within the sequence.
To prevent overfitting, the model uses a Dropout layer immediately after this step, helping to improve generalization. Next, the architecture includes a Bidirectional LSTM layer, enabling the model to consider both past and future contexts of the image sequence, better capturing the temporal relationships between frames.
An Attention layer is applied afterward, allowing the model to perform weighting of the most relevant images within the sequence, giving more importance to certain frames to improve the analysis accuracy. The outputs of this layer are passed through several Dense layers, each followed by a new Dropout layer to ensure regularization of the model. Finally, the model includes a dense layer with two neurons, responsible for binary classification.
Figure 3 shows the diagram of the model architecture, illustrating the layers nd the data flow throughout the process.

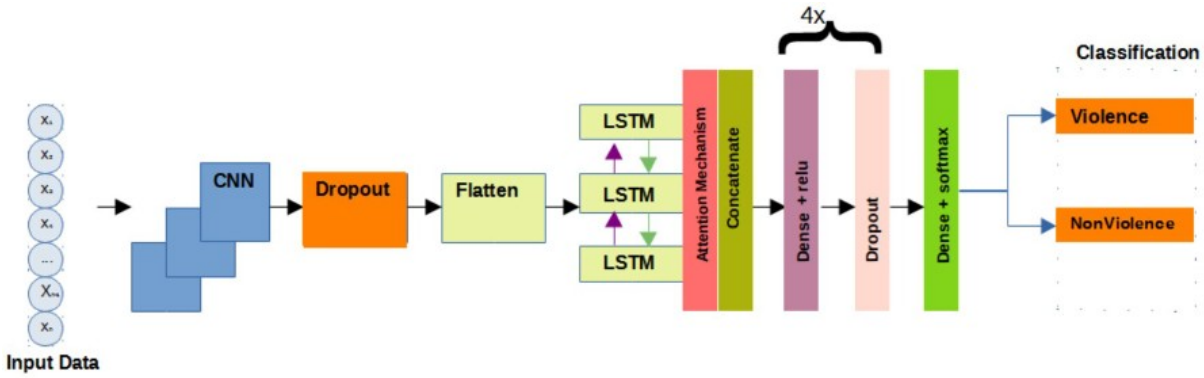


Figure 3. Architecture implemented for the experiments.

This combination of convolutional, recurrent, and attention layers allows the model to extract and learn complex, dynamic information from the images and their temporal sequences, providing a robust approach for the classification task.

3. Results and Discussion

The experiments were performed with the *MobileNetV2, DenseNet121 and InceptionV3*. It was also used *BiLSTM* in conjunction with the models. The experiments were per-

formed with and without attention mechanism. Regarding the parameters, there were variations in the minimum learning rate (*min_lr*) and batch size (*batch_size*).

Tabel 1. Training performed details with models MobileNetV2, DenseNet121 and InceptionV3.

ID	Model	Attention	min_lr	batch_size	Accuracy (%)
1	MobileNetV2	No	0.0005	128	94.25
2	DenseNet121	No	0.0005	128	94.75
3	InceptionV3	No	0.0005	128	94.25
4	MobileNetV2	No	0.00005	64	89.00
5	DenseNet121	No	0.00005	64	93.75
6	InceptionV3	No	0.00005	64	91.00
7	MobileNetV2	Yes	0.0005	128	93.25
8	DenseNet121	Yes	0.0005	128	92.50
9	InceptionV3	Yes	0.0005	128	91.75
10	MobileNetV2	Yes	0.00005	64	96.50
11	DenseNet121	Yes	0.00005	64	95.50
12	InceptionV3	Yes	0.00005	64	94.25

In experiments, the variable of interest was the accuracy obtained during training, which varied according to the use of the attention mechanism, the minimum learning rate and the size of the lot. In relation to the experiments without Attention Mechanism, in the experiment with the minimum learning rate of 0.0005 and batch size 128, the *DenseNet121* model obtained the highest accuracy, with 94.75%, followed by *MobileNetV2* with 94.25% and *InceptionV3* with 94.25%. When the minimum learning rate was reduced to 0.00005 and the batch size was adjusted to 64, *MobileNetV2* had the lowest accuracy among all experiments at 89.00%, while *DenseNet121* had a slight drop in the value of accuracy, with 93.75% and *InceptionV3* had an accuracy of 91.00%. Using *Attention Mechanism*, the models showed a slight drop in accuracy with the minimum learning rate of 0.0005 and lot size 128. The accuracy of *DenseNet121* was 93.25%, *DenseNet121* It was 92.50%, and that of *InceptionV3* was 91.75%.

Tabel 2. Best results of model performance metrics *MobileNetV2* , *DenseNet121* and *InceptionV3*.

Model	Accuracy (%)	F1-Score (Class 0)	F1-Score (Class 1)
MobileNetV2	96.50	96.00	97.00
DenseNet121	95.50	95.00	96.00
InceptionV3	94.25	94.00	94.00

When the minimum learning rate was reduced to 0.00005 and the batch size was adjusted to 64, the performance was superior compared to the no-attention experiments. The MobileNetV2 model achieved the best accuracy, with 96.50%, followed by DenseNet121 with 95.50%, and InceptionV3 with 94.25%. In Table 2, the best accuracies of each model are presented, and all models have good results with the accuracy and F1-Score performance metrics.

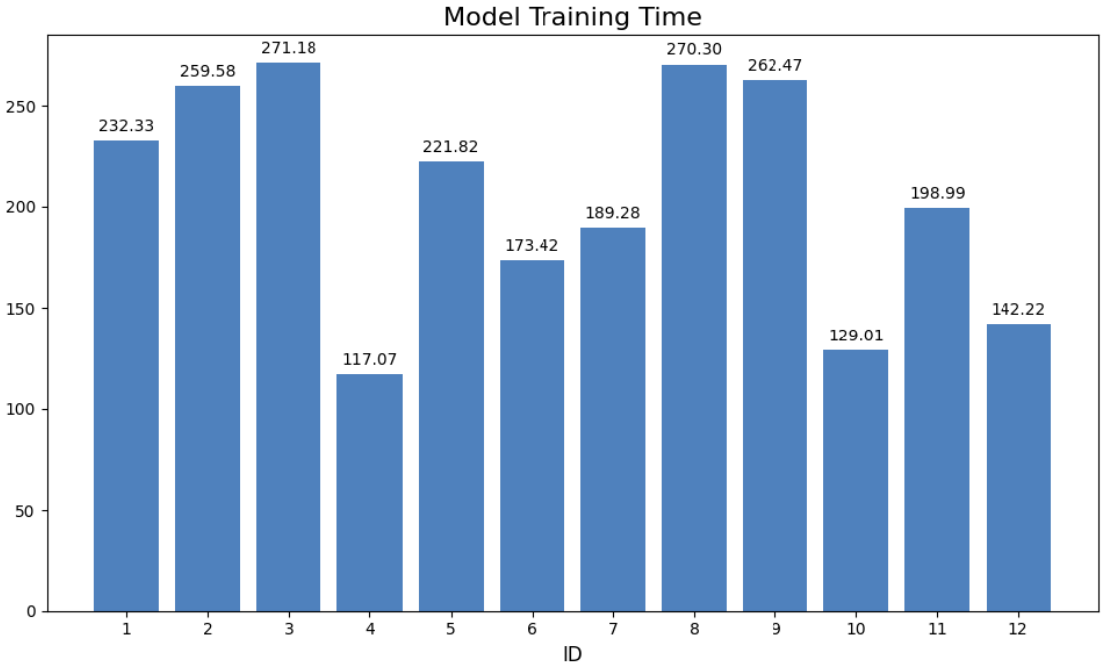
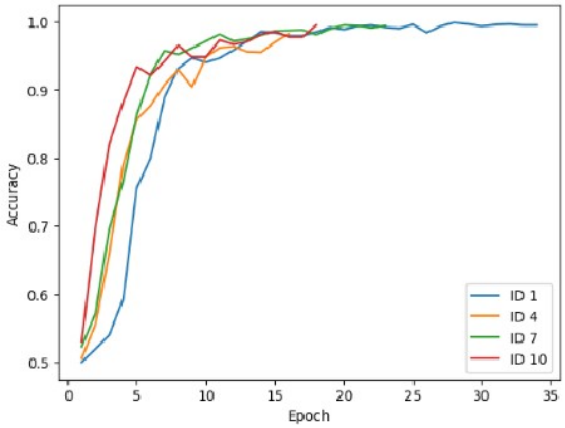
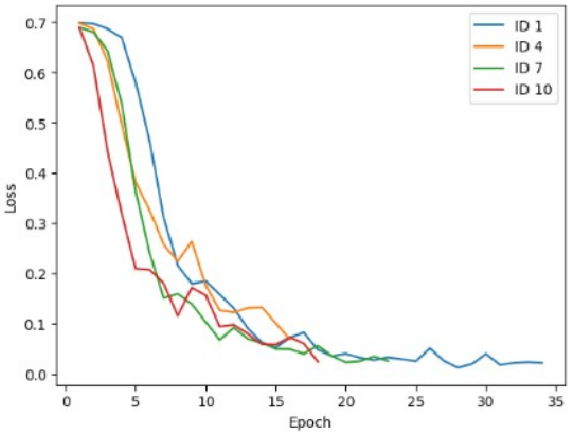


Figure 4. Training time of all experiments.

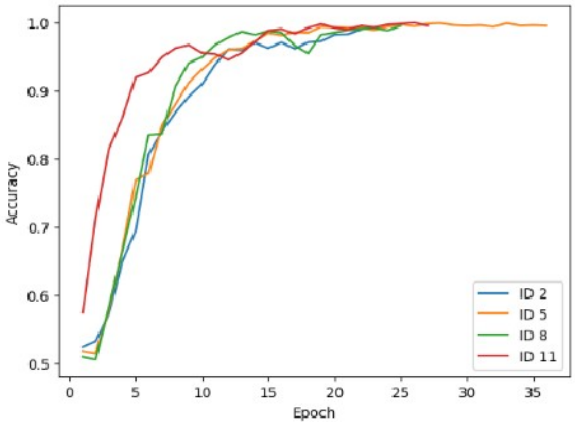
According to the Figure 4, and the reference of the experiments in the Table 1, it is observed that Attention Mechanism did not extended training time. For example, experiments 1,2,3,7,8 and 9 have the same parameter settings. Experiments ID, 1,2 and 3 do not have Attention Mechanism and experiments 7,8 and 9 have Attention Mechanism. In other words, the mechanism has no influence on training time. Another relevant analysis that can be observed in Figure 4 is that training time decreases in experiments with reduced batch sizes.



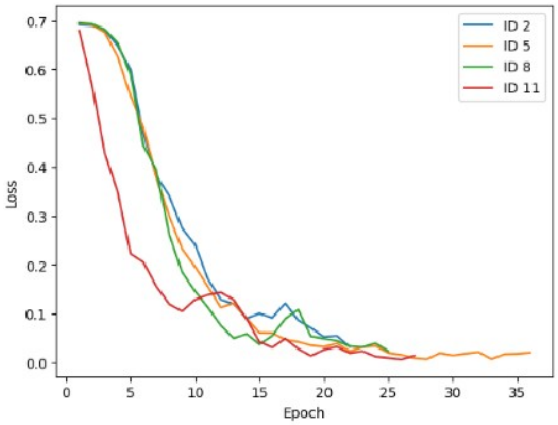
(a) Accuracy in relation to the epochs of all training of the model MobileNetV2



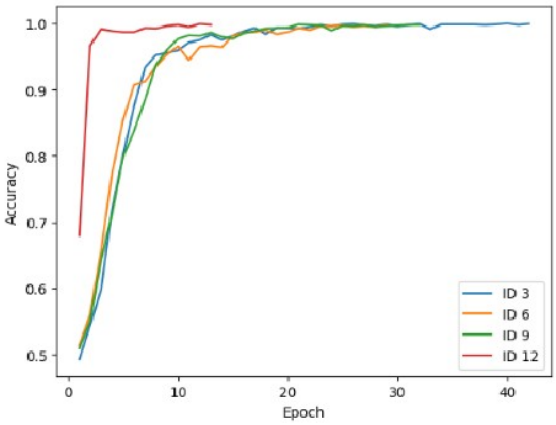
(b) Error in relation to the epochs of all training of the model MobileNetV2



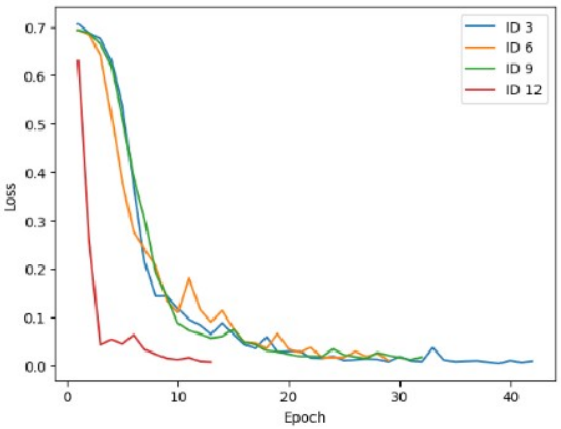
(a) Accuracy in relation to the epochs of all training of the model DenseNet121



(b) Error in relation to the epochs of all training of the model DenseNet121



(a) Accuracy in relation to the epochs of all training of the model InceptionV3



(b) Error in relation to the epochs of all training of the model InceptionV3

Figure 5. Training time of all experiments performed

In relation to the accuracy graph in Figure 5, knowing that experiment 4 does not use the mechanism, it completes training faster, in relation to the other experiments. Another relevant point in the graph of Figure 5, is that experiment 1 also does not use the mechanism and delay to finish the training. This indicates that the application Attention Mechanism has no influence on model training time. It can be observed that in all the error charts in the Figure 5 the reduced batch experiments with the mechanism have found the best minimum. These experiments sought the fastest minimums in the first times, as most of the batches gradients went to a specific direction. Another pattern of experiments 10, 11 and 12. It was that, due to the reduction of the steering speed to the minimum, the randomness led to directions in which the error increased slightly to later find better minimums.

4. Conclusion and Future Works

This study analyzed Deep Learning models for violence detection in videos, comparing MobileNetV2, DenseNet121, and InceptionV3. MobileNetV2 achieved the highest accuracy (96.50%), especially with reduced batch size and an attention mechanism. Parameter selection, such as learning rate and batch size, was crucial for optimizing performance. The Attention Mechanism showed mixed results but was beneficial in specific configurations. Future work could explore multimodal data, integrating audio, image, and sensor information for more robust scene analysis. Expanding dataset diversity and testing additional architectures may further improve results.

Acknowledgments

The authors would like to thank SAMSUNG ELETRÔNICA DA AMAZÔNIA LTDA. through the Projeto Aranouá and the Coordenação de Aperfeiçoamento de Pessoal de N´ıvel Superior (CAPES) - Programa de Excelência Acadêmica (PROEX) - Brasil for Financial Support. The present work is the result of the Research and Development (R&D) project 001/2021, signed with Instituto Federal do Amazonas and FAEPI, Brazil, which has funding from Samsung.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Proceedings International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Dashdamirov, D. (2024). Comparative analysis: Violence recognition from videos using computer vision. 2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT).

Datta, A., Shah, M., and da Vitoria Lobo, N. (2002). Person-on-person violence detection in video data. Object recognition supported by user interaction for service robots, 1:433–438 vol.1.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8):1735–1780.

Kim, J. and Grauman, K. (2009). Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In 2009 IEEE conference on computer vision and pattern recognition, pages 2921–2928. IEEE.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature, 521:436–44.

Mahadevan, V., Li, W.-X., Bhalodia, V., and Vasconcelos, N. (2010). Anomaly detection in crowded scenes. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1975–1981.

Nguyen, N. T., Phung, D. Q., Venkatesh, S., and Bui, H. H. (2005). Learning and detecting activities from movement trajectories using the hierarchical hidden markov model. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2:955–960 vol. 2.

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks. Raihan, A. S. and Ahmed, I. (2023). A bi-lstm autoencoder framework for anomaly detection – a case study of a wind power dataset.

Shenfield, A. and Howarth, M. (2020). A novel deep learning model for the detection and identification of rolling element-bearing faults. Sensors, 20(18):5112. Published: 8 September 2020.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556.

Szeliski, R. (2010). Computer Vision: Algorithms and Applications. Springer-Verlag, Berlin, Heidelberg, 1st edition.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Neural Information Processing Systems.